DOCUMENT RESUME

ED 161 676 SE 024 990

AUTHOR Gullickson, Arlen R.; Welch, Wayne W.
TITLE Applying Experimental Designs to Large-Scale Program Evaluation. Research Paper No. 2.
INSTITUTION Minnesota Univ., Minneapolis. Coll. of Education.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE [75]
GRANT NSF-GW-6800
NOTE 17p.; For related documents, see SE 024 989-999 and ED 148 632-640; Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Educational Assessment; Educational Research; *Evaluation Methods; *Program Evaluation; Research Design; *Research Methodology; *Science Education; *Statistical Analysis
IDENTIFIERS *Minnesota Research & Evaluation Project; National Science Foundation; Research Reports

ABSTRACT

This paper discusses how an experimental design can be applied to a large-scale evaluation. The purpose of the study described is to assess the success of the National Science Foundation (NSF) in achieving its goal for five comprehensive projects. This paper is divided into four main sections: (1) design; (2) sampling procedures; (3) data gathering techniques; and (4) conclusion. The majority of the paper is given to a description of the sampling and data gathering techniques used. Throughout, an emphasis is placed on procedures used and decisions made together with the reasoning behind the decisions made. Actions taken which reduced the experimental design power and advantages of using the design are discussed. It is concluded that the experimental design deserves to be considered in the evaluator's tools to be used in improving both the decision-making process and the resulting decisions. (Author/HM)

RESEARCH PAPER #2

Applying Experimental Designs to Large-Scale Program Evaluation

Arlen R. Gullickson and Wayne W. Welch

This study was supported by grant GW-6800 from the National Science Foundation to the University of Minnesota. Wayne W. Welch, Project Director.

APPLYING EXPERIMENTAL DESIGNS TO LARGE-SCALE PROGRAM EVALUATION

Arlen R. Gullickson and Wayne W. Welch

University of Minnesota

The evaluator of educational programs typically has an educational research background. That is his training has included a study of and the application of experimental design to education problems. As a result an experimental design framework is generally applied to most large evaluations. Unfortunately, in most circumstances only a modified version of a true experimental design can be used. Consequently, whether or not an experimental design can be successfully applied to an evaluation is dependent both on the evaluation problem, and upon the manner in which the experimental design is applied.

Campbell and Stanley (1) provide much help to a person selecting an experimental design, but almost no literature is available to the evaluator regarding what problems to expect in applying an experimental design. Each evaluation situation is of course unique, but many problems and decisions are common to most evaluations. Consequently, it appears important that those problems be discussed so that persons contemplating a large-scale evaluation will be cognizant of both the existing problems and some procedures that have provided workable solutions to those problems. To that end, one phase of a single evaluation is described here.

The paper is divided into four main sections: Design, Sampling Procedures, Data Gathering Techniques, and Conclusion. The majority of the paper is given to a description of the sampling and data gathering techniques used. Throughout, an emphasis is placed on procedures used and

decisions made together with the reasoning behind the decisions made. It is our hope that such a description will serve other evaluators in identifying and coping with the difficult task of evaluating large educational programs.

Design

During the summer of 1971 a proposal was written and funded (3) to evaluate a National Science Foundation (NSF) goal for its comprehensive projects, "to help schools, through the education of their instructional, resource and supervisory personnel in developing their capacity for self-improvement in science and mathematics education." The success in achieving that goal was to be assessed for five comprehensive projects. Two of the projects, at Notre Dame University and San Jose State College, were funded in mathematics; the other three projects at the Universities of Wyoming, South Dakota, and Mississippi, were funded in science. Each was to define a geographical region about its University within which it planned to achieve the desired goal.

The proposed evaluation stipulated the use of a nonequivalent pretest and posttest control group design (1). The design was suggested in order to focus on the expected change in the project region populations over a four-year period. In addition, each project was expected to use pretest information in designing the programs it intended to implement in its region.

Using the quasi-experimental design as a framework, three major tasks were undertaken during the first year: (1) instruments were developed for testing purposes, (2) people were selected for participation in the study, and (3) information was collected from participants. Only the sampling and data collection procedures are reported here.

Sampling Procedures

In order to sample, decisions were necessary regarding:

1. What should constitute the experimental unit, and from whom or what should data be collected?
2. The advisability of using a factorial design, and if used, what variables should be stratified and how should they be stratified?
3. What should constitute a control sample?
4. What should be the size of the project and control samples?
5. What techniques should be applied in selecting the sample?

The five issues are treated individually below.

Experimental unit. In an experiment results apply directly to the units or subjects observed, and from those results inferences are made to other similar units. Since the stated NSF goal is specifically "to help schools . . ." inferences from the test results would be properly directed toward schools. Consequently, even though most project money and time was to be spent with teachers, the experimental design was built with the school as the basic experimental unit or primary sampling unit (psu) as it is sometimes called.

Factorial design. Perusal of individual project proposals, meetings with the project directors and discussions with the evaluation team's advisory committee established that the projects and regions were so dissimilar that results could not be generalized from one region to another. However, in spite of their differences two important characteristics were common to all projects. Each project built its program with a subject matter orientation, and all projects made a distinction between rural and

urban areas in their efforts to involve schools. So that the two variables would be given careful attention in the analysis, both were included as factors in each region's factorial design. The urban-rural variable was handled by stratifying schools on the basis of the population of the city within which the school functioned. Four strata were formed and to optimize generalizability, the strata definitions were kept uniform across all regions. To deal with the particular subject matter emphases of a project, schools were blocked into subject matter categories. That way information peculiar to a subject could be obtained for individual schools. For example, in the South Dakota region three categories of science were formed; junior high science, biology, and chemistry.

Control regions. The question of what to use as a control for the project region schools had no completely satisfactory answer because control schools could not be selected in a way that would make them only randomly different from project region schools. In order to minimize the variance between control groups and project groups, control regions were formed from similar geographic and demographic areas. In every case the chosen control region bounded the project region on one or more sides. Certainly this method produced a set of nonequivalent control groups, but it did provide a reasonable basis for judging project successes.

Many evaluators feel the inclusion of such a control group is a luxury, and money and energy could best be spent elsewhere. Yet the statement, "In particular it should be recognized that the addition of even an unmatched or nonequivalent control group reduces greatly the equivocability of interpretation over what is obtained in Design 2, the One-Group Pretest-Posttest Design," (1:47) seems to be as timely for evaluators as for researchers.

Sample size. From an experimental viewpoint, several contributing elements made determining a precise estimate of the proper sample size a difficult task. First, a great number of dependent variables were being measured. Second, precise estimates of the variability of the dependent variables for the population being studied were not available. Third, no project region had specified the degree of intended impact on its region during the four-year tenure of the Comprehensive Grant.

Given the above obstacles, an estimate of desired sample size was made using the following rationale. If the projects do have a meaningful effect on their regions, that effect should cause at least one-half a standard deviation change in the dependent variables over a four-year period. And, if an ANOVA test were used to detect that pre-post change, the sample size should be large enough to reduce the risk of Type 1 and Type 2 errors to the following levels: a 5 percent chance of not detecting the change in the event it did occur, and a 10 percent chance of incorrectly stating that the change had occurred in the event it did not.

Use of those parameters together with a procedure and power function graph for analysis of variance (2) resulted in a sample size estimate of 35. Because pre-post differences were to be detected in each subject strata, 35 then was the desired number per subject area.

As is explained later, data from each of the regions were to be collected through the use of regional meetings. It was important for comprehensive project development reasons to have a large representative group of schools attend these meetings (approximately 100 from each region). Three subject strata in science suggested 105 participants per science region while the two strata (junior high school and senior high school) in mathematics

suggested a sample size of 70. Because of the need to conduct several studies across science and math, we decided to make the samples approximately equal. Accordingly, our target samples were 105 schools per science region, and 100 schools per math region. Because of anticipated nonresponse, we oversampled in each region by approximately 40 percent.

Because of the expense involved with the large-scale regional meetings, the project's financial status required that the control samples be about one-half the project region samples. Consequently, 75 and 70 schools were sampled in each of the respective science and mathematics control regions.

Sampling procedure. A simple, efficient, systematic sampling method was used to select 1,095 schools (730 in the experimental group and 365 in the control) in the project and control regions.

To sample systematically, all N primary sampling units (psu's) in the population are listed. The number N is then divided by the number, n, of psu's desired in the sample. The number, X, obtained by that division is used as follows: The first psu to be included in the sample is randomly selected from the list; the second psu to be selected is X units down the list from the first; the third psu is 2X units down the list from the first; the fourth is 3X units down the list from the first, and so on with the final psu selected being (n-1)X units down the list from the first. By this method, having selected the first unit, all other psu's to be included in the sample are automatically determined.

For the NSF evaluation, an alphabetical list of schools (psu's) was developed for each stratum in each region. The general systematic sampling process described above was then carried out for all strata in every region. Although not a true random method, systematic sampling did produce a sample considered sufficiently random for the evaluation.

The sampling of subjects within each of the schools proved to be more difficult to handle and resulted in a less satisfactory solution. Practical considerations would not allow many sources within each school to be tapped, so the school principal, one teacher, and one class were chosen to represent each school.

The school principal from each sampled school was invited to participate and was asked to select one teacher as a co-participant. In selecting a teacher, each principal was asked to follow two criteria. First, the teacher was to be chosen from among all teachers of a particular subject (the subject depended upon which cell of the factorial design the school was in). Second, the names of all eligible teachers were to be placed in a hat and one name, the participant's, was to be drawn from it.

The classes which participated also were to meet two criteria: (1) the class was to be taught by the teacher that participated and (2) the class was to be selected by a random procedure spelled out in the written directions included with the packet of instruments sent to each teacher.

Obviously the selection criteria for teachers and classes could be violated if the principals and teachers chose, but there seemed to be no economical alternative to the steps taken.

Data Gathering Techniques

Project regions. Meetings were held within each region to obtain project region data. While attending a meeting, principals and teachers completed biographic questionnaires, attitude, and achievement measures and each participating teacher was to take a test packet back to his school and administer it to the selected class. One factor contributing to the decision

to hold regional meetings was a desire by the evaluation team to include a teacher achievement and a teaching knowledge measure as part of the evaluation. The National Teachers Exam (NTE) was judged the best instrument for that purpose. It requires two hours to administer, and is "secure" i.e., it can be administered only under strict supervision.

A second factor contributing to the need for the regional meetings was the large block of time needed by participants for completing evaluation instruments. The time required was estimated at one hour for principals and three hours for teachers. That seemed an inordinate amount of time to ask a principal or teacher to spend responding to a series of instruments distributed through the mails.

A third and deciding factor was that meetings would be used to the benefit of the NSF projects and the NSF program in general. It was hoped that through participation, principals and teachers would become familiar with the NSF and its projects. At the same time, the meetings would help NSF program directors to determine the needs of each region and plan better ways to make their resources beneficial to the schools.

In late January, letters inviting the principal and his randomly selected teacher to participate in the regional meeting were sent to each of the 730 selected project region schools. Each was a personal letter to the school's principal, on NSF stationery, and hand-signed by the Director of the NSF Academic Year Study Program. Because of the nature and status of NSF, it was expected that virtually all principals would return the enclosed postcard and that approximately 70 percent would agree to participate. On February 14, the Monday following the February 11 response deadline, it was clear that the evaluation team's expectations had been unrealistically high. For example,

only 40 percent of the principals in the Notre Dame University region responded, and in no region did more than two-thirds respond.

The large number of nonrespondents placed the regional meeting concept in jeopardy. Therein lay a serious problem. Should additional schools be sampled with the hope of obtaining enough participants from an additional sample to ensure 100-105 participating schools within each project region? Or should a maximum effort be made to obtain sufficient participants from the schools already sampled? The possibility of sampling additional schools was finally rejected as an alternative since: (1) a large new sample would be required for some regions; (2) a second batch of initial letters would reach principals only a short time before some meetings were to take place, and would probably result in an even smaller acceptance rate than had occurred for the original sample; and (3) having only a small percent of the sample participate raises questions about the "representativeness" of the participating sample. Sampling additional schools would not have improved the representativeness problem; in fact it would probably have been detrimental.

The representativeness question in particular led to the decision to make a concerted attempt to get a large percent of the nonrespondents to respond affirmatively to the request. Each principal who did not respond to the original letter was sent a follow-up letter. Telephone calls were then made to all principals who either did not respond to the second letter, or indicated on a return postcard that they were unsure about attending. Altogether approximately 200 principals were called, the need for their participation in a regional meeting explained to them, and a verbal request for their participation made. As Table 1 shows, the follow-up procedure greatly increased the total number of schools agreeing to participate.

Insert Table 1 about here

Control regions. Unlike the data collection method used for the project region sample, all instruments were mailed to control school participants. Due to the high cost of the NTE, it was administered only to project sample teachers. With that obstacle of mailing questionnaires removed, there remained no compelling reason to hold meetings, and the difference in cost was sufficient reason for change. The same contact procedure used for the project regions was used for the control, except that no telephone calls were made for the control groups. Overall, the control schools responded better to the evaluation than had the regional schools. Table 2 gives a summary of the control region sampling response. (Because a large percent of both regional and control schools invited did not agree to participate, a separate study was undertaken to compare selected characteristics of participating schools and nonparticipating schools.)

Insert Table 2 about here

One additional feature of the control region procedure merits mentioning. Since the control sample teachers and principals had virtually nothing to gain by participating, it was felt that a large percentage of principals contacted would choose not to involve themselves or others in their school. To improve the situation, it was suggested that an incentive of $5 for each principal and teacher be offered. Because such an incentive might have no effect or even be detrimental rather than beneficial, and because the same general evaluation plan was to be used again for posttesting purposes, the effect of an incentive in increasing participation was tested. The control sample was randomly split;

TABLE 1

PROJECT REGION SAMPLING RESPONSE

Region	Schools Agreeing to Participate			
	Original Request		After Follow-up Procedure	
	Number	Percent of Total Sampled	Number	Percent of Total Sampled
Notre Dame	37	26	62	44
San Jose	49	35	97	69
Mississippi	46	31	66	44
South Dakota	72	48	96	64
Wyoming	76	51	99	66

$\bar{X} = 57\%$

TABLE 2

CONTROL REGION SAMPLING RESPONSE

Region	Schools Agreeing to Participate	
	Number	Percent of Total Sampled
Notre Dame Control	47	67
San Jose Control	41	59
Mississippi Control	40	53
South Dakota Control	56	75
Wyoming Control	50	67

$\bar{x} = 64\%$

half of the principals received letters containing no incentive and the other half received a letter offering the aforementioned $5 incentive. The study is still in progress and its results are unknown at this time.

Conclusion

A review of several points may serve to highlight the concerns evidenced here, and simultaneously clarify what we consider to be the proper role of experimental design in evaluation. It is clear that when both evaluative functions and design functions could not be served equally well, most decisions were made from an evaluation basis. It seems obvious that if experimental design is to serve evaluation needs, that will always be the case.

The following are actions taken in the NSF evaluation which reduced the experimental design's power. A quasi-experimental rather than a true-experimental design was used. The selected sample did not have the size desired for detecting differences between the project regions and control regions. Each factorial design originally incorporated four city strata. However, so few schools were included from very large cities that the four strata were collapsed to two, in order that the data analysis would have reasonable power for detecting the project impact on schools in cities of different sizes. In addition, the techniques used in selecting teachers and classes, and the difference in methods used for collecting the data from project and control regions may pose serious problems in the interpretation of data if those methods appear to be the cause of differential results among the schools.

Though some expectations were reduced by the compromises, the design still offers a credible basis from which to make evaluative decisions and

discuss results. The quasi-experimental design provided for pre-post comparisons of each project region with its control group. The factorial design gave focus to stated curriculum emphases, and allowed for differentiating each project's impact on rural and urban areas. Design considerations caused an awareness of critical concerns in selecting the control group, determining the sample size, and in distributing the sample within the factorial design. Also, design considerations provided an atmosphere that fostered the growth of research (the incentive study) compatible to evaluation concerns. Most important, the design established alternatives and caused the evaluation team to set priorities based on expectations of each alternative's value for the evaluation and for the NSF Comprehensive Projects as a whole.

Ideally an experimental design can be as potent for the evaluator as the researchers. Practically, based on experiences such as the evaluation just described, we know that experimental design can at a minimum serve in defining strategies and setting priorities. Therefore, it deserves to be included in the evaluator's "tool kit" to be used in improving both the decision-making process and the resulting decisions.

REFERENCES

1. Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

2. Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, California: Wadsworth, 1968.

3. Welch, W. W. Proposal for a research grant to study the design, implementation, and efficacy of the NSF comprehensive programs for teacher education. (Grant No. GW 6800) Washington, D.C., National Science Foundation, 1971.